A Brief Review of TDT Methods and a Perspective on Its Power

Jing Hua Zhao

Section of Genetic Epidemiology/Biostatistics

Department of Psychological Medicine

KCL Insititute of Psychiatry

j.zhao@iop.kcl.ac.uk

www.iop.kcl.ac.uk/IoP/Departments/PsychMed/GEpiBSt/index.stm

May 29, 2002

Outline

- A short literature review
- Some work on power (1999, personal communication to Dr Long, UC Davis)
- Software
- Discussion

Literature on TDT

- GRR, HHRR (Falk-Rubinstein, Terwilliger-Ott, Spielman, Schaid-Sommer-Knapp), see, e.g. Schaid and Sommer (1994)
- Multiallelic (Bickeboller, Rice, Spielman, Sham-Curtis, Schaid)
- Sibling controls (Curtis, Beohnke-Langefield, Spielman, Horvath/Siegmund)
- Quantitative traits (Allison, Rabinowitz, Fulker)
- Resources
  - Section website (1,600 references in Endnote)
  - linkage.rockefeller.edu (by year, author)
  - Schaid (1998), Schaid & Rowland (1998), Schaid (1999a, 1999b), Schaid & Rowland (2000), Knapp (1999a, 1999b, 1999c), Horvath et al. (2000) give more accounts about TDT. A more recent summary of the statistical issues in TDT was given by Zhao (2000).

A perspective on power of TDT

Beside work of Risch & Merikangas (1996), the power of TDT using both family trios and sibling controls has been investigated by Kaplan et al. (1997), Baur & Knapp (1997), Knapp (1999a, 1999b), Cervino & Hill (2000), Monks et al. (1998). Others included Baur & Knapp (1997), Trégouët at al. (2001).

The work of Risch & Merikangas (1996) has provoked discussions concerning the use of linkage versus association design (Scott et al. 1997; Müller-Myhsok & Abel 1997; Long et al. 1997; Risch & Teng 1998; Morton & Collins 1998; Teng & Risch 1999; Long & Langley 1999; Camp 1997; Knapp 1999a). These are particularly relevant in the context of increasingly dense genetic map. When serving as a guideline for designing a TDT study, it is unfortunate that the paper of Risch & Merikangas (1996) contained a programming error. A correction will shed light on later work such as Camp (1997, 1999) and represent the original conclusions in terms of sibling instead of genotypic relative risks. Comparison with that of a case-control design will further clarify important issues made by those discussants.

Some preparations

- A disease susceptibility locus has two alleles A and a
- Allele frequencies p and $q = 1 - p$
- Penetrances f_0, f_1 and f_2
- Recombination fraction between disease locus and marker θ

General model

- $K = p^2 f^2 + 2pqf_1 + q^2 f_0$
- $V_A = 2pq[p(f_2 - f_1) + q(f1 - f_0)]^2$
- $V_D = p^2 q^2 (f_2 - 2f_1 + f_0)^2$
- Let $\Psi = \theta^2 + (1-\theta)^2$, then IBD probabilities for affected sib-pair are as follows

$$
\begin{aligned}
P(IBD = 0) &= \frac{1}{4} - \frac{(\Psi - .5)V_A + (2\Psi - \Psi^2 - .75)V_D}{4(K^2 + .5V_A + .25V_D)} \\
P(IBD = 1) &= \frac{1}{2} - \frac{2(\Psi^2 - \Psi + .25)V_D}{4(K^2 + .5V_A + .25V_D)} \\
P(IBD = 2) &= \frac{1}{4} + \frac{(\Psi - .5)V_A + (\Psi^2 - .25)V_D}{4(K^2 + .5V_A + .25V_D)}
\end{aligned}
$$

  Under no linkage, affected sib pair sharing 0, 1, and 2 alleles IBD 1/4, 1/2, and 1/4 (Suarez et al. 1978)

- Offspring and sibling RRs are $\lambda_O = 1 + 0.5V_A/K^2$ and $\lambda_S = 1 + (0.5V_A + 0.25V_D)/K^2$
- The probabilities of siblings sharing none or one allele by descent are $z_0 = 0.25/\lambda_S$ and $z_1 = 0.5\lambda_O/\lambda_S$, and the nonshared probability is $Y = 1 - 0.5z_1 - z_0$

Multiplicative Model

- GRRs for AA, Aa, aa are γ^2, γ and 1.
- $K = p^2\gamma^2 + 2pq\gamma + q^2 = (p\gamma + q)^2$, $V_A = 2pq(\gamma - 1)^2(p\gamma + q)^2$ and $V_D = p^2q^2(\gamma - 1)^4$.
- $\lambda_O = 1 + w$ and $\lambda_S = (1 + 0.5w)^2$, where $w = pq(\gamma - 1)^2/(p\gamma + q)^2$.
- $Y = 1 - 0.25(\lambda_O + 1)/\lambda_S = (1 + w)/(2 + w)$.
- $h = P$(a parent being heterozygous (H) | an affected child)= $P(H)P(\text{AffChild}|H)/P(\text{AffChild})$, or $2pq(0.5p(\gamma^2 + \gamma) + 0.5q(\gamma + 1))/(p\gamma + q)^2$.

Test statistic I

- N independent identically distributed random variables B_i with mean 0 and variance 1 under the null hypothesis, mean mean μ and variance σ^2 under the alternative hypothesis.
- Statistic $\Sigma_{i=1}^{N} B_i/N$ has mean 0 and variance 1 under the null but mean $\sqrt{N}\mu$ and variance σ^2 under the alternative.
- Sample size N for a given significance level α and power $1-\beta$ can be estimated by $(Z_\alpha - \sigma Z_{1-\beta})^2/\mu^2$.

Test statistic II

- ASP: the allele shared and nonshared from the ith parent is a r.v. denoted by B_i and scored 1 and -1. Under the null hypothesis, the shared and nonshared each has probability 0.5 so the mean and variance of B_i are 0 and 1. Under the alternative $\mu = 2Y - 1$ and $\sigma^2 = 4Y(1-Y)$. Assuming sharing of alleles from both parent to be independent, the required sample size for affected sib-pair under $\theta = 0$ and no linkage disequilibrium is $N = (z_\alpha - \sigma z_{1-\beta})^2/2\mu^2$, $Y = (1+w)/(2+w)$.
- TDT with ASP: $h = P($ a parent of an ASP is heterozygous $) = pq(\gamma+1)^2/[2(\gamma p+q)^2 + pq(\gamma-1)^2]$, the same sample size formula applies but the required number of families is half the expected number since there are two independent affected sibs.
- Singleton: $B_i = 1/\sqrt{h}, h = pq(\gamma+1)/(p\gamma+q)$ if parent is heterozygous and transmits A, $B_i = 0$ if parent is homozygous, $B_i = -1/\sqrt{h}$ if parent is heterozygous and transmits a. Under the null hypothesis the mean and variance of B_i are 0 and 1, whereas under the alternative they are $\sqrt{h}(\gamma-1)/(\gamma+1)$ and $1 - h[(\gamma-1)/(\gamma+1)]^2$, respectively.

Comparisons with case-control design

- Consider a statistic directly testing association between marker and disease (Long et al. 1997).
- With a randomly ascertained population sample, under HWE multiplicative model in which the genotypic relative risk is γ, the frequencies of the three disease genotypes AA, Aa and aa in cases are $\pi\gamma^2$, $2\pi\gamma pq$, and πq^2, respectively, where π is the "baseline" probability that an individual with aa genotype being affected. Similarly the three frequencies in controls are $(1-\pi\gamma^2)p^2$, $2(1-\pi\gamma)pq$ and $(1-\pi)q^2$. A unit χ^2 statistic can be constructed using the following 2×2 table:

	affected genotype	nonaffected genotype	
A	$\pi\gamma^2 p^2 + \pi\gamma pq$	$(1-\pi\gamma^2)p^2 + (1-\pi\gamma)pq$	p
a	$\pi\gamma pq + \pi q^2$	$(1-\pi\gamma)pq + (1-\pi)q^2$	q
	$\pi(\gamma p + q)^2$	$1 - \pi(\gamma p + q)^2$	1

or equivalently

	affected genotype	nonaffected genotype	
A	$\pi\gamma p(\gamma p+q)$	$p-\pi\gamma p(\gamma p+q)$	p
a	$\pi q(\gamma p+q)$	$q-\pi q(\gamma p+q)$	q
	$\pi(\gamma p+q)^2$	$1-\pi(\gamma p+q)^2$	1

So that the expected unit frequencies (E) are $\pi p(\gamma p+q)^2$, $\pi q(\gamma p+q)^2$, $p-\pi p(\gamma p+q)^2$ and $q-\pi q(\gamma p+q)^2$, the discrepancies between observed and expected frequencies all have factor $\pi pq(\gamma p+q)(\gamma-1)$ but with negative sign before the second and the third items. The statistic $\chi^2=\Sigma(O-E)^2/E$ is then easy to obtain. Under the null hypothesis $\gamma=1$, χ^2 has central χ^2 distribution, whereas under the alternative $\gamma>1$ and χ^2 follows noncentral $\chi^2_{1,\delta}$ distribution with noncentrality parameter $\delta=[\pi pq(\gamma-1)^2]/[1-\pi(\gamma p+q)^2]$ or $\delta=[\gamma^2 p+q-(\gamma p+q)^2]/[1-\pi(\gamma p+q)^2]$ It is equivalently to derive the power by $Y=\sqrt{X^2}\sim N(\sqrt{\delta},1)$. $1-\beta=\Phi(-Z<Y<Z)$, where Z is a preassigned standard normal deviate.

TDT and ASP designs

- Let the genome-wide significance level and type II error rate be $\alpha = 5 \times 10^{-8}$ and $\beta = 0.2$
- Table 1. Column N_{asp} corrects the calculation from the original paper (as $N_{asp}(\times)$). The Alzheimer's disease model is based on Scott et al. (1997).
- It turns out that with $\gamma \leq 2$, the expected marker-sharing only marginally exceeds 50% for any allele frequency (p). The use of linkage would need practically nonachievable sample size. Nonetheless, direct tests of association with a disease locus itself can still be quite strong. But it may involve large amount of statistical testing of associated alleles.

Table 1: Comparison of linkage and association in nuclear families required for identification of disease gene

		Linkage			Association				
γ	p	Y	N_L	P_A	Het_1	N_{tdt}	Het_2	$N_{asp/tdt}$	$N_{asp}(\times)$
4.0	0.01	0.520	6400	0.800	0.048	1098	0.112	235	4260
	0.10	0.597	276	0.800	0.346	150	0.537	48	185
	0.50	0.576	445	0.800	0.500	103	0.424	61	297
	0.80	0.529	3022	0.800	0.235	222	0.163	161	2013
2.0	0.01	0.502	445835	0.667	0.029	5823	0.043	1970	296710
	0.10	0.518	8085	0.667	0.245	695	0.323	264	5382
	0.50	0.526	3751	0.667	0.500	340	0.474	180	2498
	0.80	0.512	17904	0.667	0.267	640	0.217	394	11917
1.5	0.01	0.501	6943229	0.600	0.025	19320	0.031	7776	4620807
	0.10	0.505	101898	0.600	0.214	2218	0.253	941	67816
	0.50	0.510	27040	0.600	0.500	949	0.490	484	17997
	0.80	0.505	101898	0.600	0.286	1663	0.253	941	67816
Alzheimer's:									
4.5	0.15	0.626	163	0.818	0.460	100	0.621	36	109

γ=genotypic risk ratio; p=frequency of disease allele A; Y=probability of allele sharing; N_L=number of ASP families required for linkage; P_A=probability of transmitting disease allele A; Het_1,Het_2=proportions of heterozygous parents; N_{tdt}=number of family trios; $N_{asp/tdt}$=number of ASP families

Case-control design

- Table 2. Long et al. (1997) used approximation $1 - \beta = \Phi(Z - \sqrt{\delta})$ by taking the area under normal curve in the lower tail as negligible. It seems this approximation is quite good (Columns 3-5).

Table 2: Estimated sample sizes required for association detection

		Long et al. (1997)			Actual calculation		
γ	p	1%	5%	10%	1%	5%	10%
4.0	0.01	46681	8959	4244	46637	8951	4240
	0.10	8180	1570	744	8172	1568	743
	0.50	10890	2090	990	10880	2088	989
	0.80	31473	6040	2861	31444	6035	2859
2.0	0.01	403970	77530	36725	403593	77457	36690
	0.10	52709	10116	4792	52660	10106	4787
	0.50	35284	6772	3208	35252	6765	3205
	0.80	79390	15236	7217	79316	15222	7211
1.5	0.01	$> 10^6$	307055	145447	1598429	306769	145312
	0.10	192104	36869	17464	191925	36834	17448
	0.50	98012	18810	8910	97921	18793	8902
	0.80	192104	36869	17464	191926	36834	17448

- Clearly it is most favourable for diseases that are relatively common, which has important implications for complex traits. When the disease is relatively common, the disease-allele frequency is intermediate and

its effect small, statistical power comparable to that of standard family-based linkage studies is achieved with a smaller number of randomly sampled individuals. While statistical power in terms of required sample size is important, practicality also needs to be considered. For efficiently diagnosed late onset disease such as non-insulin-dependent diabetes and hypertension, it may not be possible to type parents for affected sibling studies. A further point is that the actual number of individuals genotyped needed would be doubled for linkage, tripled for singleton, and quadrupled for sib-pair, assuming both parents are genotyped in a affected offspring study (Long et al. 1997). Long et al. (1997) noted when there is actual marker-disease data, a Fisher's exact test can be used in the case of population sample to detect association.

- Case-control design is commonly believed to be more powerful assuming there is no population stratification (Morton & Collins 1998; Devlin & Roeder 1999; Bacanu et al. 2000; Risch 2000; Pritchard & Rosenberg 1999; Pritchard et al. 2000).

- Seltman et al. (2001) proposed to associate the evolutionary tree with TDT.

Generalisations

- One criticism was that genotypic relative risk γ rather than sibling relative risk λ_s was employed–since even with large γ, λ_s could still be small, this was remedied in Risch (1997). Camp (1997, 1999) and Knapp (1999a).

- Incomplete disequilibrium (Müller-Myhsok & Abel 1997) and the same genetic model and one biallelic marker with alleles B and b of frequencies m and $1-m$, they considered the following table,

	B	b	
A	$pm+\delta$	$p(1-m)-\delta$	p
a	$(1-p)m-\delta$	$(1-p)(1-m)+\delta$	$1-p$
	m	$1-m$	1

  where the parameter $\delta = P(AB) - pm$ achieves its maximum when $P(AB)$ is $\min(m, p)$.

- Let $\alpha_1 = P(A|B) = (pm+\delta)/m = p + \delta/m$ and $\alpha_2 = P(A|b) = [p(1-m)-\delta]/(1-m) = p - \delta/(1-m)$, for TDT with trios the probability that Bb parent transmits B to the affected child is $\tau(B) = P(\text{aff}|B)/[P(\text{aff}|B) + P(\text{aff}|b)]$, the prior

probabilities of transmitting B and b are both 0.5, and $P(\text{aff}|B) = [\gamma\alpha_1 + (1-\alpha_1)]D$ and $P(\text{aff}|b) = [\gamma\alpha_2 + (1-\alpha_2)]D$, where D is the probability that is a subject is affected given he carries allele a. It turns that out $\tau(B) = [1+(\gamma-1)\alpha_1]/[2+(\gamma-1)(\alpha_1+\alpha_2)]$, which only reduces to $\gamma/(1+\gamma)$ when $m = p$ whereas when m/p departs from unity the power diminishes substantially. Abel & Müller-Myhsok (1998) also expressed TDT as $\Lambda = 2n\ln[q\ln(q)+(1-q)\ln(1-q)-\ln(0.5)]$, n being the number of heterozygous parents, q being the probability Bb transmitting B, and considered the difference between Λ and the classic TDT $d(q) = 2n[q\ln(q)+(1-q)\ln(1-q)-\ln(0.5)-2(q-0.5)^2]$.

- Assuming that under the null hypothesis $q \sim N(\tau(B), \sigma^2)$, $\sigma^2 = \tau(B)(1-\tau(B))/n$, the number of heterozygous parents is obtained by solving

$$2n[q_\beta \ln(q_\beta) + (1-q_\beta)\ln(1-q_\beta) - \ln(0.5)] = Z_\alpha^2$$

  where $q_\beta = \tau(B) - \sigma Z_{1-\beta}$.

- From n, the required number of families $N = n/2h$ can be estimated, h is the probability that a parent with an affected child is heterozygous given by Risch & Merikangas (1996), or by Müller-Myhsok & Abel

(1997) as $u/[u + m^2[1 + (\gamma - 1)\alpha_1] + (1 - m)^2[1 + (\gamma - 1)\alpha_2]$ with $u = m(1-m)[2+(\gamma-1)(\alpha_1+\alpha_2)]$.

- Setting $\gamma = 4$, $\alpha = 5 \times 10^{-8}$ and $1 - \beta = 0.80$, for the maximum likelihood test of $p = 0.5$ (ML-TDT), $N = 139$ for $p = m = 0.10$ and $N = 96$ for $p = m = 0.50$, compared with 150 and 103 in Table 1.

Relation between case-control and TDT

- The magnitude of disease-marker association detected by TDT can be estimated by binomial proportion $T = M_1/(M_1 + M_2)$, for the number of times high (M_1) and low (M_2) risk alleles are transmitted, respectively. Let m=allele frequencies of high risk marker allele, g_i=P(genotype i|aff)P(aff)/P(genotype i). This statistic can be rewritten as a function of the relative frequencies of the marker genotypes among the affected offspring of a heterozygous parent as follows,

Child G	Allele passed	P(G)	P(aff|G)	Frequencies
M_1M_1	M_1	$0.5m$	g_2	$0.5m(g_2)$
M_1M_2	M_1	$0.5(1-m)$	g_1	$0.5(1-m)(g_1)$
M_1M_2	M_2	$0.5m$	g_1	$0.5m(g_1)$
M_2M_2	M_2	$0.5(1-m)$	g_0	$0.5(1-m)(g_0)$

- let $A = 0.5m(g_2) + 0.5(1-m)(g_1)$, $B = 0.5m(g_1) + 0.5(1-m)(g_0)$, then $T = A/(A+B)$, T could be referred to power/sample size formula for proportion.

Examples

Mitchell (2000) gave the following examples,

Disease	Marker	T[case-control, TDT (C.I.)]
Cleft lip/palate	TGFA	0.67 0.77 (0.63-0.91)
Cleft palate	MSX1	0.59 0.68 (0.51-0.85)
Spina bifida	5,10 mthfr	0.60 0.56 (0.49-0.63)
IDDM	5'FP	0.77 0.63 (0.54-0.72)

- The estimate of T is outside confidence interval (C.I.) predicted by TDT only in the last example.
- For the first disorder, T_1=0.62, T_0=0.5, type I error α=0.05, type II error β=0.20, then the required sample size becomes $N = [1.64[0.5(1-0.5)]^{0.5} + 0.84[0.62(1-0.62)]^{0.5}]^2/(0.62-0.50)^2 \approx 105$, or N=266 for α=0.001.

Other work

- McGinnis (1998, 2000) further clarified the relative power of TDT and affected sib-pair method, including expression of general mode of inheritance.
- Cervino & Hill (2000) compared TRANSMIT, SIBASSOC/STDT and RCTDT (Knapp 1999b) in a variety of scenarios. From their simulation, the classic "likelihood-ratio association test" would result in high type I error when there is a population substructure suggesting simple parameterisation in terms of transmission and different allele frequencies may not be enough to characterise the effect of substructure whereas introducing more parameters would increase the degrees of freedom and reduce power. It seems TRANSMIT is quite robust to substructure and provides both correct type I error and reasonable power. Since TRANSMIT is based on a multiplicative model it is also most powerful when this is true. RCTDT is also appealing.

Power Calculation, which version to use?

- Knapp et al. (1995) on Schaid and Sommer (1993)
- The use of a variance formula in Risch and Merikangas (1996) was incorrect but corrected in 1997, the assumption of complete LD and greater discrepancies between disease and marker allele frequencies were questioned by Müller-Myhsok and Abel (1997)
- The independence assumption of parental transmission is correct with multiplicative models but wrong for general models in Camp (1997), corrected in 1999
- Knapp (1999a) provided another correction of Camp (1997), appropriate for general mode of inheritance and consistent with Baur and Knapp (1997)

Software

- AFBAC
- ASPEX
- ETDT, CETDT, TDTmax (logistic regression/Bradley-Terry model), MASC
- FBAT (XDT)
- GASSOC (Dom, Rec)
- GENEHUNTER
- PDT
- QTDT
- TRANSMIT

HHRR and TRANSMIT

- Single biallelic marker
- Single multiallelic marker (data from Homero)

  Haplotype-based score tests

  1 df tests for individual haplotypes

H	O	E	Untrans'd	Var(O-E)	χ^2_1
1	5	6.7688	8.5376	2.7187	1.1508
2	13	13.344	13.689	4.9193	0.024096
3	71	60.305	49.61	23.19	4.9325
4	172	165.98	159.96	48.989	0.73992
5	385	392.76	400.53	91.915	0.65574
6	372	365.24	358.48	95.226	0.47961
7	204	212.2	220.4	60.388	1.114
8	92	98.131	104.26	29.581	1.2708
9	14	13.264	12.528	5.5605	0.097398

  Global chisquared test, on 8 degrees of freedom = 9.8164

  HHRR statistic = 7.2114 (suggested by Dr Andrew Makoff)

- Haplotypes?